Explain various stages in DataStage

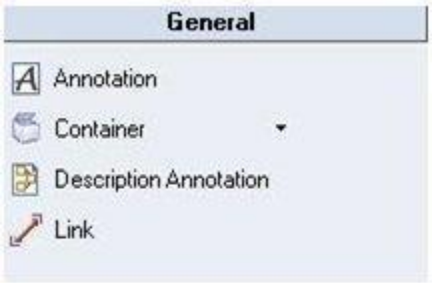
DataStage is an ETL tool that extracts, transforms, loads data from source to target. DataStage promotes market research by offering reliable data for business intelligence. There are several stages in DataStage I listed them out in this article.

If you want to Gain In-depth Knowledge on **DataStage**, please go through this link **DataStage Training**
DataStage and QualityStage stages sortes into the following logical sections:

- General elements
- Quality Development and Debug Stages
- Database stages
- Real- stages
- Debug and development stages
- Sequence activities

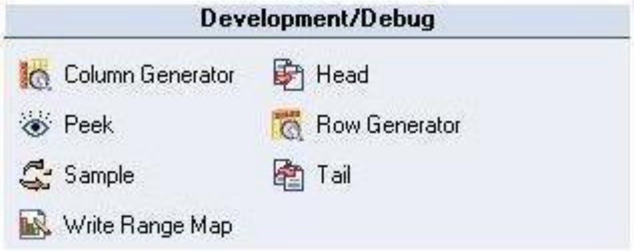
The common and famous stages used in DataStage and QualityStage are discussed below and specific important features are explained below

General Elements in DataStage:



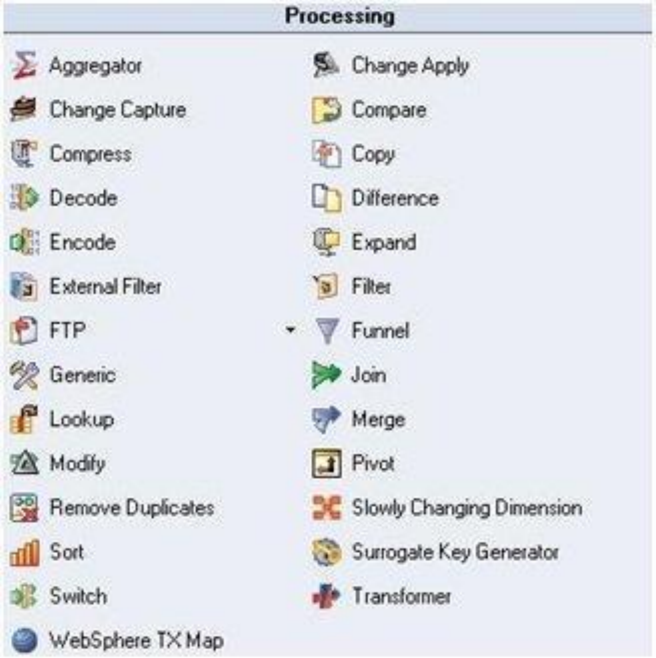
- A Link indicates the data flow. The main types of links are reference, stream, and lookup stage in DataStage.
- The **container** can be shared or private; the main function is to make it visually clear so that the complicated job design of the database allows for ease of use and identification of the system.
- In theory, **annotation** used to apply floating DataStage explanations and annotations to work posts. This offers a way to chart the ETL cycle and is used to explain the job in question.

2. Debug and development in DataStage:



- **Row generator** replicates a collection of data that sticks to the appropriate metadata. It is useful for creation and exploration.
- **Column generator** may add one or more columns to the flow, and may generate column test data.
- **Peek** : The values of the column are registered, and displayed in the manager. It can have multiple links to output, but a single connection to input.
- **Sample** is operating on a set of input data and has two modes:- percent mode and time mode.
- **Head**: At each partition, Head occupies first N rows and copies them from the input to the output data set. You can chose last N rows from each of the partitions.
- **Write Range Map**: Writes a type where the range partitioning method makes the dataset accessible.

3. Processing Stages in DataStage:

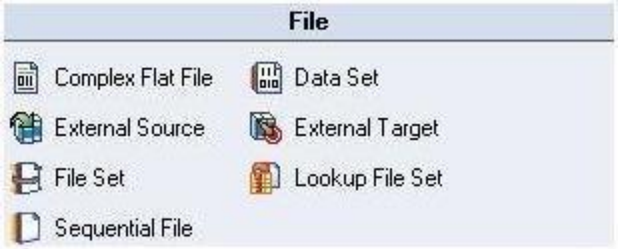


IBM Infosphere DataStage consists of individual stages connected together. It defines data flow from data source to data target. Typically a stage has at least one input and/or data output. Some stages, however, can accept more than one input and output to more than one stage.

- **Aggregator** combines data vertically by grouping incoming data streams and calculating summaries for and group (sum, count, min, cap, variance, etc.). Two methods can group the data: hash table or pre-sort.
- **Copy-**copy input data (a single stream) to one or more output data flows FTP stage uses FTP protocol to pass data to a remote computer Filter filters out records that do not meet specified requirements.
- **Funnel** blends several sources.
- **Join** blends two or more key-column values inputs. Close principle to the DBMS SQL link. Can have one left and multiple right inputs and generate a single output stream (no reject link).
- **Lookup** blends two or more key-column values inputs. Lookup stage can have one source and several tables. Records need not sort and produce a single output stream and reject connection.
- **Merge** blends one master input with several update inputs by main column values. All inputs sort, and multiple reject connections will catch unknown secondary entries.
- **Modify** stage changes the dataset record scheme. Useful for column renaming, non-default data type conversions, and null handling Eliminate duplicates stage includes a single sorted data set as input.
- **Transformer** stage manages extracted data, performs data validation, transformations, and lookups.
- **Change Capture:** Describes a single set of records representing changes made before and after two sets of input data and output.
- **Adjust Apply**: Make changes to a collection of predata to assess a collection of post-data.It carries out a record-by-record analysis of two sets of input data and generates a single set of data whose records reflect the difference between them.
- **Checksum**: Generates and adds checksum from the stated columns in a row. We use it to assess variations between documents.
- **Compare** conducts a column-by-column analysis of two presorted input data sets. It has two input links and one output link.
- Encode data encoding commands such as gzip.
- Decode a data set previously encoded with Encode Point.
- **External Filter** allows defining an operating system command that acts as a filter on the processed generic data level, enabling users to call an OSH operator from the DataStage level with options as required.
- **Pivot** for horizontal pivoting. It maps multiple columns in a single column in multiple output rows. Pivoting data results in a dataset with fewer columns, but more rows.
- Surrogate Key Generator produces column surrogate key and manages key source.
- Switch stage assigns each input row to a selector field value-based output relation. It provides a similar principle to the transition in most programming languages.

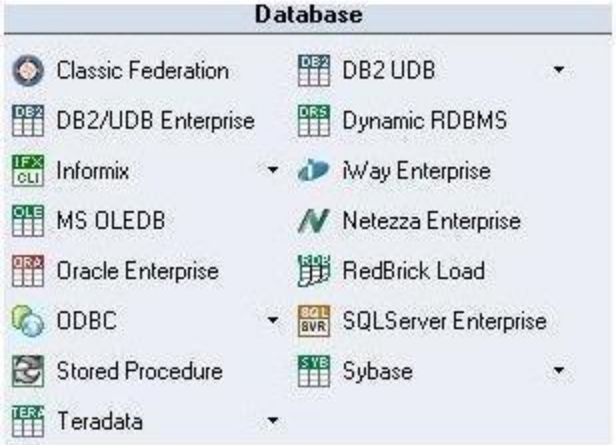
- **Compress:** Loads a data set using a GZIP utility (or LINUX / UNIX compress command) Expand transforms a previously compressed data set back into raw binary data.

4. File stagesin DataStage:



- The sequential file is to access or write data from one or more flat (sequential) files.
- Stage Data Set allows users to access data from or write data to a series of files. File sets are OS files each with an extension of.ds and one or more data files.
- File Set stage allows users to read or write data from a series of files. Unlike datasets, file sets maintain formatting and are readable by other applications.
- The compact flat file allows the reading of organized files from compact file systems on a mainframe computer similar to a header, MVS data sets, and truck.
- External Source lets generate data from multiple source programs.
- Lookup File Set is identical to Stage set file. Lookups will use partitioned hashed format.

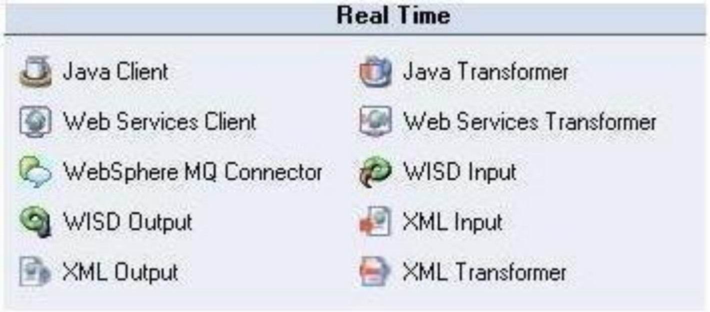
5. Database Stages in DataStage:



- Oracle Enterprise allows data reading into an Oracle database.
- ODBC Enterprise allows data retrieved from and written to a database called an ODBC source. It's used in Microsoft Access and Excel spreadsheets for data processing.

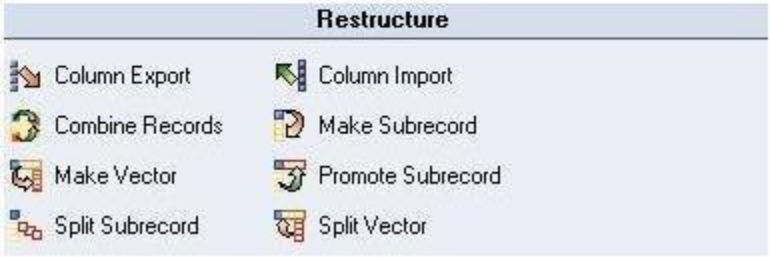
- DB2/UDB Enterprise allows a DB2 Server to write, access and access data.
- Teradata lets you write, read data to a warehouse of Teradata data. There are three Teradata stages; Teradata, Business, and Multiload connector.
- SQL Server Enterprise lets the Microsoft SQLI Server 2005 and 2008 database write and read data.
- Sybase allows data to be read and written into Sybase Databases.
- DB2, Oracle, Sybase, Teradata, and Microsoft SQL Server are stored process stage chains.
- MS OLEDB is used to retrieve information from all forms of sources of information, such as an ISAM server, a relational source, or a table.
- Dynamic Relational Stage uses interfaces such as Microsoft SQL, DB2, Oracle, Sybase and Informix to read and write from and to a specific supported relational DB.
- DB2 UDB (API or Load) 6, Informix (CLI or Load). Real Time Stages m XML Input stage allows for the conversion of hierarchical XML data into flat relational data sets.

6. Real Time Stages in DataStage:



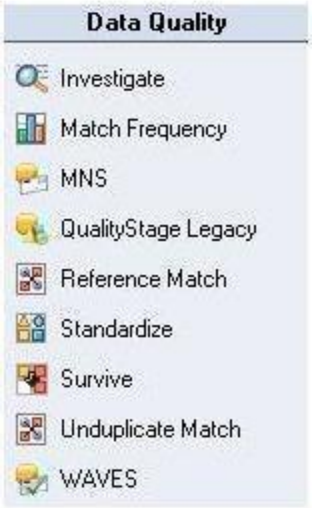
- XML Input Stage enables translating hierarchical XML data to flat relational data sets.
- XML Output for XML structures.
- XML Transformer transforms stylesheet XML documents.
- Websphere MQ stages have a networking menu to access IBM WebSphere MQ enterprise messaging systems.
- Java client stage may also be useful as a goal and lookup. The kit includes 3 groups i.e input, output and reject.

7. Restructure stages in DataStage:



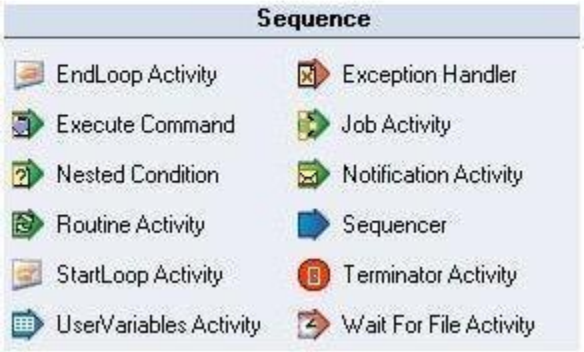
- Export column, exporting data to a single binary column or string from multiple columns of different data types. It can have one output, input and reject connection.
- **Column import:** Corresponding to Column Export, characteristically used to divide inward bound data into various columns.
- Combine records stage community rows with exact same keys, over sub-record vectors.
- Let sub-records combine different input vectors into a sub-record vector whose columns have the same names as the original vectors.
- Make vector combines defined input vectors into a column vector Promote sub-records to columns at the top point.
- Split sub-records split an input sub-records area into a series of top-level columns.

8. Data QualityStage stages in DataStage:



- Investigate stage predicts data module of suitable columns of all the source file information. Offers methods for exploring word and character.
- Match frequency stage gets feedback from a computer, database, or processing stage and produces a distribution report for an event.
- MNS: Refers to WAVES international address standardization: refers to the verification and improvement method of addresses around the world.

Sequence type stages in DataStage:



Operation shows DataStage server or related function to execute.

Notification Process used to transfer DataStage emails to client-described recipients.

Sequencer used to synchronize control flow of various acts in job progression.

Terminator operation allows to shut down all progress until those conditions continue.

Wait for file Operation waits for an correct file to appear or fade away, and launches dispensation.

Conclusion:

I hope you understood about several stages in DataStage. You can learn more from **DataStage Online Training**.